Internet Engineering Task Force (IETF)
Request for Comments: 7829
Category: Standards Track
ISSN: 2070-1721

Y. Nishida
GE Global Research
P. Natarajan
Cisco Systems
A. Caro
BBN Technologies
P. Amer
University of Delaware
K. Nielsen
Ericsson

April 2016

SCTP-PF: A Quick Failover Algorithm for the Stream Control Transmission Protocol

Abstract

The Stream Control Transmission Protocol (SCTP) supports multihoming. However, when the failover operation specified in RFC 4960 is followed, there can be significant delay and performance degradation in the data transfer path failover. This document specifies a quick failover algorithm and introduces the SCTP Potentially Failed (SCTP-PF) destination state in SCTP Path Management.

This document also specifies a dormant state operation of SCTP that is required to be followed by an SCTP-PF implementation, but it may equally well be applied by a standard SCTP implementation, as described in RFC 4960.

Additionally, this document introduces an alternative switchback operation mode called "Primary Path Switchover" that will be beneficial in certain situations. This mode of operation applies to both a standard SCTP implementation and an SCTP-PF implementation.

The procedures defined in the document require only minimal modifications to the specification in RFC 4960. The procedures are sender-side only and do not impact the SCTP receiver.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7829.

Copyright Notice

Table of Contents

1. Introduction

The Stream Control Transmission Protocol (SCTP) specified in [RFC4960] supports multihoming at the transport layer. SCTP's multihoming features include failure detection and failover procedures to provide network interface redundancy and improved end-to-end fault tolerance. In SCTP's current failure detection procedure, the sender must experience Path.Max.Retrans (PMR) number of consecutive failed timer-based retransmissions on a destination address before detecting a path failure. Until detecting the path failure, the sender continues to transmit data on the failed path. The prolonged time in which SCTP as described in [RFC4960] continues to use a failed path severely degrades the performance of the protocol. To address this problem, this document specifies a quick failover algorithm called "SCTP-PF" based on the introduction of a new Potentially Failed (PF) path state in SCTP path management. The

performance deficiencies of the failover operation described in RFC 4960, and the improvements obtainable from the introduction of a PF state in SCTP, were proposed and documented in [NATARAJAN09] for Concurrent Multipath Transfer SCTP [IYENGAR06].

While SCTP-PF can accelerate the failover process and improve performance, the risk that an SCTP endpoint might enter the dormant state where all destination addresses are inactive can be increased. [RFC4960] leaves the protocol operation during dormant state to implementations and encourages avoiding entering the state as much as possible by careful tuning of the PMR and Association.Max.Retrans (AMR) parameters. We specify a dormant state operation for SCTP-PF, which makes SCTP-PF provide the same disruption tolerance as [RFC4960] despite the fact that the dormant state may be entered more quickly. The dormant state operation may equally well be applied by an implementation of [RFC4960] and will serve here to provide added fault tolerance for situations where the tuning of the PMR and AMR parameters fail to provide adequate prevention of the entering of the dormant state.

The operation after the recovery of a failed path also impacts the performance of the protocol. With the procedures specified in [RFC4960], SCTP will (after a failover from the primary path) switch back to use the primary path for data transfer as soon as this path becomes available again. From a performance perspective, such a forced switchback of the data transmission path can be suboptimal as the Congestion Window (CWND) towards the original primary destination address has to be rebuilt once data transfer resumes, [CARO02]. As an optional alternative to the switchback operation of [RFC4960], this document specifies an alternative Primary Path Switchover procedure that avoids such forced switchbacks of the data transfer path. The Primary Path Switchover operation was originally proposed in [CARO02].

While SCTP-PF is primarily motivated by a desire to improve the multihomed operation, the feature also applies to SCTP single-homed operation. Here the algorithm serves to provide increased failure detection on idle associations, whereas the failover or switchback aspects of the algorithm will not be activated. This is discussed in more detail in Appendix C.

A brief description of the motivation for the introduction of the PF state, including a discussion of alternative approaches to mitigate the deficiencies of the failover operation in [RFC4960], are given in the appendices. Discussion of path-bouncing effects that might be caused by frequent switchovers are also provided there.

2. Conventions and Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. SCTP with Potentially Failed (SCTP-PF) Destination State

3.1. Overview

To minimize the performance impact during failover, the sender should avoid transmitting data to a failed destination address as early as possible. In the SCTP path management scheme described in [RFC4960], the sender stops transmitting data to a destination address only after the destination address is marked inactive. This process takes a significant amount of time as it requires the error counter of the destination address to exceed the PMR threshold. The issue cannot simply be mitigated by lowering the PMR threshold because this may result in spurious failure detection and unnecessary prevention of the usage of a preferred primary path. Also, due to the coupled tuning of the PMR and the AMR parameter values in [RFC4960], lowering the PMR threshold may result in lowering the AMR threshold, which would result in a decrease of the fault tolerance of SCTP.

The solution provided in this document is to extend the SCTP path management scheme of [RFC4960] by the addition of the PF state as an intermediate state in between the active and inactive state of a destination address in the path management scheme of [RFC4960], and let the failover of data transfer away from a destination address be driven by the entering of the PF state instead of by the entering of the inactive state. Thereby, SCTP may perform quick failover without negatively impacting the overall fault tolerance of SCTP as described in [RFC4960]. At the same time, HEARTBEAT probing based on Retransmission Timeout (RTO) is initiated towards a destination address once it enters PF state. Thereby, SCTP may quickly ascertain whether network connectivity towards the destination address is broken or whether the failover was spurious. In the case where the failover was spurious, data transfer may quickly resume towards the original destination address.

The new failure detection algorithm assumes that loss detected by a timeout implies either severe congestion or network connectivity failure. It recommends that, by default, a destination address be classified as PF at the occurrence of the first timeout.

3.2. Specification of the SCTP-PF Procedures

The SCTP-PF operation is specified as follows:

1. The sender maintains a new tunable SCTP Protocol Parameter called PotentiallyFailed.Max.Retrans (PFMR). The PFMR defines the new intermediate PF threshold on the destination address error counter. When this threshold is exceeded, the destination address is classified as PF. The RECOMMENDED value of PFMR is 0. If PFMR is set to be greater than or equal to PMR, the resulting PF threshold will be so high that the destination address will reach the inactive state before it can be classified as PF.

2. The error counter of an active destination address is incremented or cleared as specified in [RFC4960]. This means that the error counter of the destination address in active state will be incremented each time the Timer T3 retransmission (T3-rtx) timer expires, or each time a HEARTBEAT chunk is sent when idle and not acknowledged within an RTO. When the value in the destination address error counter exceeds PFMR, the endpoint MUST mark the destination address as in the PF state.

3. An SCTP-PF sender SHOULD NOT send data to destination addresses in PF state when alternative destination addresses in active state are available. Specifically, this means that:

   i. When there is outbound data to send and the destination address presently used for data transmission is in PF state, the sender SHOULD choose a destination address in active state, if one exists, and use this destination address for data transmission.

   ii. As specified in Section 6.4.1 of [RFC4960], when the sender retransmits data that has timed out, they should attempt to pick a new destination address for data retransmission. In this case, the sender SHOULD choose an alternate destination transport address in active state, if one exists.

   iii. When there is outbound data to send and the SCTP user explicitly requests to send data to a destination address in PF state, the sender SHOULD send the data to an alternate destination address in active state if one exists.

When choosing among multiple destination addresses in active state, an SCTP sender will follow the guiding principles of Section 6.4.1 of [RFC4960] by choosing the most divergent source-destination pairs compared with, for (the aforementioned points i and ii):

i. the destination address in PF state that it performs a failover from, and

ii. the destination address towards which the data timed out.

Rules for picking the most divergent source-destination pair are an implementation decision and are not specified within this document.

In all cases, the sender MUST NOT change the state of the chosen destination address, whether this state be active or PF, and it MUST NOT clear the error counter of the destination address as a result of choosing the destination address for data transmission.

4. When the destination addresses are all in PF state, or some are in PF state and some in inactive state, the sender MUST choose one destination address in PF state and SHOULD transmit or retransmit data to this destination address using the following rules:

   i. The sender SHOULD choose the destination in PF state with the lowest error count (fewest consecutive timeouts) for data transmission and transmit or retransmit data to this destination.

   ii. When there are multiple destination addresses in PF state with same error count, the sender should let the choice among the multiple destination addresses in PF state with equal error count be based on the principles of choosing the most divergent source-destination pairs when executing (potentially consecutive) retransmission outlined in Section 6.4.1 of [RFC4960]. Rules for picking the most divergent source-destination pairs are an implementation decision and are not specified within this document.

   The sender MUST NOT change the state and the error counter of any destination addresses as the result of the selection.

5. The HB.Interval of the Path Heartbeat function of [RFC4960] MUST be ignored for destination addresses in PF state. Instead, HEARTBEAT chunks are sent to destination addresses in PF state

once per RTO. HEARTBEAT chunks SHOULD be sent to destination addresses in PF state, but the sending of HEARTBEATs MUST honor whether or not the Path Heartbeat function (Section 8.3 of [RFC4960]) is enabled for the destination address. That is, if the Path Heartbeat function is disabled for the destination address in question, HEARTBEATs MUST NOT be sent. Note that when the Path Heartbeat function is disabled, it may take longer to transition a destination address in PF state back to active state.

6. HEARTBEATs are sent when a destination address reaches the PF state. When a HEARTBEAT chunk is not acknowledged within the RTO, the sender increments the error counter and exponentially backs off the RTO value. If the error counter is less than PMR, the sender transmits another packet containing the HEARTBEAT chunk immediately after timeout expiration on the previous HEARTBEAT. When data is being transmitted to a destination address in the PF state, the transmission of a HEARTBEAT chunk MAY be omitted in the case where the receipt of a Selective Acknowledgment (SACK) of the data or a T3-rtx timer expiration on the data can provide equivalent information, such as the case where the data chunk has been transmitted to a single destination address only. Likewise, the timeout of a HEARTBEAT chunk MAY be ignored if data is outstanding towards the destination address.

7. When the sender receives a HEARTBEAT ACK from a HEARTBEAT sent to a destination address in PF state, the sender SHOULD clear the error counter of the destination address and transition the destination address back to active state. However, there may be a situation where HEARTBEAT chunks can go through while DATA chunks cannot. Hence, in a situation where a HEARTBEAT ACK arrives while there is data outstanding towards the destination address to which the HEARTBEAT was sent, then an implementation MAY choose to not have the HEARTBEAT ACK reset the error counter, but have the error counter reset await the fate of the outstanding data transmission. This situation can happen when data is sent to a destination address in PF state. When the sender resumes data transmission on a destination address after a transition of the destination address from PF to active state, it MUST do this following the prescriptions of Section 7.2 of [RFC4960].

8. Additional PMR - PFMR consecutive timeouts on a destination address in PF state confirm the path failure, upon which the destination address transitions to the inactive state. As described in [RFC4960], the sender SHOULD (i) notify the Upper Layer Protocol (ULP) about this state transition, and (ii)

transmit HEARTBEAT chunks to the inactive destination address at a lower HB.Interval frequency as described in Section 8.3 of [RFC4960] (when the Path Heartbeat function is enabled for the destination address).

9. Acknowledgments for chunks that have been transmitted to multiple destinations (i.e., a chunk that has been retransmitted to a different destination address than the destination address to which the chunk was first transmitted) SHOULD NOT clear the error count for an inactive destination address and SHOULD NOT move a destination address in PF state back to active state, since a sender cannot disambiguate whether the ACK was for the original transmission or the retransmission(s). An SCTP sender MAY clear the error counter and move a destination address back to active state by information other than acknowledgments, when it can uniquely determine which destination, among multiple destination addresses, the chunk reached. This document makes no reference to what such information could consist of, nor how such information could be obtained.

10. Acknowledgments for data chunks that have been transmitted to one destination address only MUST clear the error counter for the destination address and MUST transition a destination address in PF state back to active state. This situation can happen when new data is sent to a destination address in the PF state. It can also happen in situations where the destination address is in the PF state due to the occurrence of a spurious T3-rtx timer and acknowledgments start to arrive for data sent prior to occurrence of the spurious T3-rtx and data has not yet been retransmitted towards other destinations. This document does not specify special handling for detection of, or reaction to, spurious T3-rtx timeouts, e.g., for special operation vis-a-vis the congestion control handling or data retransmission operation towards a destination address that undergoes a transition from active to PF to active state due to a spurious T3-rtx timeout. But it is noted that this is an area that would benefit from additional attention, experimentation, and specification for single-homed SCTP as well as for multihomed SCTP protocol operation.

11. When all destination addresses are in inactive state, and SCTP protocol operation thus is said to be in dormant state, the prescriptions given in Section 4 shall be followed.

12. The SCTP stack SHOULD expose the PF state of its destination addresses to the ULP as well as provide the means to notify the ULP of state transitions of its destination addresses from active to PF, and vice versa. However, it is recommended that

an SCTP stack implementing SCTP-PF also allows for the ULP to be kept ignorant of the PF state of its destinations and the associated state transitions, thus allowing for retention of the simpler state transition model of [RFC4960] in the ULP. For this reason, it is recommended that an SCTP stack implementing SCTP-PF also provide the ULP with the means to suppress exposure of the PF state and the associated state transitions.

4. Dormant State Operation

In a situation with complete disruption of the communication in between the SCTP endpoints, the aggressive HEARTBEAT transmissions of SCTP-PF on destination addresses in PF state may make the association enter dormant state faster than a standard SCTP implementation of [RFC4960] given the same setting of PMR and AMR. For example, an SCTP association with two destination addresses would typically reach dormant state in half the time of an SCTP implementation of [RFC4960] in such situations. This is because an SCTP PF sender will send HEARTBEATs and data retransmissions in parallel with RTO intervals when there are multiple destinations addresses in PF state. This argument presumes that RTO << HB.Interval of [RFC4960]. With the design goal that SCTP-PF shall provide the same level of disruption tolerance as a standard SCTP implementation with the same PMR and AMR setting, we prescribe that an SCTP-PF implementation SHOULD operate as described in Section 4.1 during dormant state.

An SCTP-PF implementation MAY choose a different dormant state operation than the one described in Section 4.1 provided that the solution chosen does not decrease the fault tolerance of the SCTP-PF operation.

The prescription below for SCTP-PF dormant state handling MUST NOT be coupled to the value of the PFMR, but solely to the activation of SCTP-PF logic in an SCTP implementation.

It is noted that the below dormant state operation can also provide enhanced disruption tolerance to a standard SCTP implementation that doesn’t support SCTP-PF. Thus, it can be sensible for a standard SCTP implementation to follow this mode of operation. For a standard SCTP implementation, the continuation of data transmission during dormant state makes the fault tolerance of SCTP be more robust towards situations where some, or all, alternative paths of an SCTP association approach, or reach, inactive state before the primary path used for data transmission observes trouble.

4.1. SCTP Dormant State Procedure

1. When the destination addresses are all in inactive state and data is available for transfer, the sender MUST choose one destination and transmit data to this destination address.

2. The sender MUST NOT change the state of the chosen destination address (it remains in inactive state) and MUST NOT clear the error counter of the destination address as a result of choosing the destination address for data transmission.

3. The sender SHOULD choose the destination in inactive state with the lowest error count (fewest consecutive timeouts) for data transmission. When there are multiple destinations with the same error count in inactive state, the sender SHOULD attempt to pick the most divergent source -- destination pair from the last source -- destination pair where failure was observed. Rules for picking the most divergent source-destination pair are an implementation decision and are not specified within this document. To support differentiation of inactive destination addresses based on their error count, SCTP will need to allow for incrementing of the destination address error counters up to some reasonable limit above PMR+1, thus changing the prescriptions of Section 8.3 of [RFC4960] in this respect. The exact limit to apply is not specified in this document, but it is considered reasonable enough to require that the limit be an order of magnitude higher than the PMR value. A sender MAY choose to deploy other strategies than the strategy defined here. The strategy to prioritize the last active destination address, i.e., the destination address with the fewest error counts is optimal when some paths are permanently inactive, but suboptimal when path instability is transient.

5. Primary Path Switchover

The objective of the Primary Path Switchover operation is to allow the SCTP sender to continue data transmission on a new working path even when the old primary destination address becomes active again. This is achieved by having SCTP perform a switchover of the primary path to the new working path if the error counter of the primary path exceeds a certain threshold. This mode of operation can be applied not only to SCTP-PF implementations, but also to implementations of [RFC4960].

The Primary Path Switchover operation requires only sender-side changes. The details are:

1. The sender maintains a new tunable parameter, called Primary.Switchover.Max.Retrans (PSMR). For SCTP-PF implementations, the PSMR MUST be set greater than or equal to the PFMR value. For implementations of [RFC4960], the PSMR MUST be set greater than or equal to the PMR value. Implementations MUST reject any other values of PSMR.

2. When the path error counter on a set primary path exceeds PSMR, the SCTP implementation MUST autonomously select and set a new primary path.

3. The primary path selected by the SCTP implementation MUST be the path that, at the given time, would be chosen for data transfer. A previously failed primary path can be used as a data transfer path as per normal path selection when the present data transfer path fails.

4. For SCTP-PF, the recommended value of PSMR is PFMR when Primary Path Switchover operation mode is used. This means that no forced switchback to a previously failed primary path is performed. An SCTP-PF implementation of Primary Path Switchover MUST support the setting of PSMR = PFMR. An SCTP-PF implementation of Primary Path Switchover MAY support setting of PSMR > PFMR.

5. For standard SCTP, the recommended value of PSMR is PMR when Primary Path Switchover is used. This means that no forced switchback to a previously failed primary path is performed. A standard SCTP implementation of Primary Path Switchover MUST support the setting of PSMR = PMR. A standard SCTP implementation of Primary Path Switchover MAY support larger settings of PSMR > PMR.

6. It MUST be possible to disable the Primary Path Switchover operation and obtain the standard switchback operation of [RFC4960].

The manner of switchover operation that is most optimal in a given scenario depends on the relative quality of a set primary path versus the quality of alternative paths available as well as on the extent to which it is desired for the mode of operation to enforce traffic distribution over a number of network paths. That is, load distribution of traffic from multiple SCTP associations may be enforced by distribution of the set primary paths with the switchback operation of [RFC4960]. However, as switchback behavior of [RFC4960]

is suboptimal in certain situations, especially in scenarios where a number of equally good paths are available, an SCTP implementation MAY support also, as alternative behavior, the Primary Path Switchover mode of operation and MAY enable it based on applications' requests.

For an SCTP implementation that implements the Primary Path Switchover operation, this specification RECOMMENDS that the standard switchback operation of [RFC4960] be retained as the default operation.

6. Suggested SCTP Protocol Parameter Values

This document does not alter the value recommendation for the SCTP Protocol Parameters defined in [RFC4960].

The following protocol parameter is RECOMMENDED:

PotentiallyFailed.Max.Retrans (PFMR) - 0

7. Socket API Considerations

This section describes how the socket API defined in [RFC6458] is extended to provide a way for the application to control and observe the SCTP-PF behavior as well as the Primary Path Switchover function.

Please note that this section is informational only.

A socket API implementation based on [RFC6458] is, by means of the existing SCTP_PEER_ADDR_CHANGE event, extended to provide the event notification when a peer address enters or leaves the PF state as well as the socket API implementation is extended to expose the PF state of a peer address in the existing SCTP_GET_PEER_ADDR_INFO structure.

Furthermore, two new read/write socket options for the level IPPROTO_SCTP and the name SCTP_PEER_ADDR_THLDS and SCTP_EXPOSE_POTENTIALLY_FAILED_STATE are defined as described below. The first socket option is used to control the values of the PFMR and PSMR parameters described in Sections 3 and 5. The second one controls the exposition of the PF path state.

Support for the SCTP_PEER_ADDR_THLDS and SCTP_EXPOSE_POTENTIALLY_FAILED_STATE socket options also needs to be added to the function sctp_opt_info().

7.1. Support for the Potentially Failed Path State

As defined in [RFC6458], the SCTP_PEER_ADDR_CHANGE event is provided if the status of a peer address changes. In addition to the state changes described in [RFC6458], this event is also provided if a peer address enters or leaves the PF state. The notification as defined in [RFC6458] uses the following structure:

```
struct sctp_paddr_change {
  uint16_t spc_type;
  uint16_t spc_flags;
  uint32_t spc_length;
  struct sockaddr_storage spc_aaddr;
  uint32_t spc_state;
  uint32_t spc_error;
  sctp_assoc_t spc_assoc_id;
}
```

[RFC6458] defines the constants SCTP_ADDR_AVAILABLE, SCTP_ADDR_UNREACHABLE, SCTP_ADDR_REMOVED, SCTP_ADDR_ADDED, and SCTP_ADDR_MADE_PRIM to be provided in the spc_state field. This document defines the new additional constant SCTP_ADDR_POTENTIALLY_FAILED, which is reported if the affected address becomes PF.

The SCTP_GET_PEER_ADDR_INFO socket option defined in [RFC6458] can be used to query the state of a peer address. It uses the following structure:

```
struct sctp_paddrinfo {
  sctp_assoc_t spinfo_assoc_id;
  struct sockaddr_storage spinfo_address;
  int32_t spinfo_state;
  uint32_t spinfo_cwnd;
  uint32_t spinfo_srtt;
  uint32_t spinfo_rto;
  uint32_t spinfo_mtu;
};
```

[RFC6458] defines the constants SCTP_UNCONFIRMED, SCTP_ACTIVE, and SCTP_INACTIVE to be provided in the spinfo_state field. This document defines the new additional constant SCTP_POTENTIALLY_FAILED, which is reported if the peer address is PF.

7.2. Peer Address Thresholds (SCTP_PEER_ADDR_THLDS) Socket Option

Applications can control the SCTP-PF behavior by getting or setting the number of consecutive timeouts before a peer address is considered PF or unreachable. The same socket option is used by applications to set and get the number of timeouts before the primary path is changed automatically by the Primary Path Switchover function. This socket option uses the level IPPROTO_SCTP and the name SCTP_PEER_ADDR_THLDS.

The following structure is used to access and modify the thresholds:

```
struct sctp_paddrthlds {
  sctp_assoc_t spt_assoc_id;
  struct sockaddr_storage spt_address;
  uint16_t spt_pathmaxrxt;
  uint16_t spt_pathpfthld;
  uint16_t spt_pathcpthld;
};
```

spt_assoc_id: This parameter is ignored for one-to-one style sockets. For one-to-many style sockets, the application may fill in an association identifier or SCTP_FUTURE_ASSOC. It is an error to use SCTP_{CURRENT|ALL}_ASSOC in spt_assoc_id.

spt_address: This specifies which peer address is of interest. If a wildcard address is provided, this socket option applies to all current and future peer addresses.

spt_pathmaxrxt: Each peer address of interest is considered unreachable, if its path error counter exceeds spt_pathmaxrxt.

spt_pathpfthld: Each peer address of interest is considered PF, if its path error counter exceeds spt_pathpfthld.

spt_pathcpthld: Each peer address of interest is not considered the primary remote address anymore, if its path error counter exceeds spt_pathcpthld. Using a value of 0xffff disables the selection of a new primary peer address. If an implementation does not support the automatic selection of a new primary address, it should indicate an error with errno set to EINVAL if a value different from 0xffff is used in spt_pathcpthld. For SCTP-PF, the setting of spt_pathcpthld < spt_pathpfthld should be rejected with errno set to EINVAL. For standard SCTP, the setting of spt_pathcpthld < spt_pathmaxrxt should be rejected with errno set to EINVAL. An SCTP-PF implementation may support only setting of spt_pathcpthld = spt_pathpfthld and spt_pathcpthld = 0xffff and a standard SCTP

implementation may support only setting of spt_pathcpthld = spt_pathmaxrxt and spt_pathcpthld = 0xffff. In these cases, SCTP shall reject setting of other values with errno set to EINVAL.

7.3. Exposing the Potentially Failed Path State (SCTP_EXPOSE_POTENTIALLY_FAILED_STATE) Socket Option

Applications can control the exposure of the PF path state in the SCTP_PEER_ADDR_CHANGE event and the SCTP_GET_PEER_ADDR_INFO as described in Section 7.1. The default value is implementation specific.

This socket option uses the level IPPROTO_SCTP and the name SCTP_EXPOSE_POTENTIALLY_FAILED_STATE.

The following structure is used to control the exposition of the PF path state:

```
struct sctp_assoc_value {
  sctp_assoc_t assoc_id;
  uint32_t assoc_value;
};
```

assoc_id: This parameter is ignored for one-to-one style sockets. For one-to-many style sockets, the application may fill in an association identifier or SCTP_FUTURE_ASSOC. It is an error to use SCTP_{CURRENT|ALL}_ASSOC in assoc_id.

assoc_value: The PF path state is exposed if, and only if, this parameter is non-zero.

8. Security Considerations

Security considerations for the use of SCTP and its APIs are discussed in [RFC4960] and [RFC6458].

The logic introduced by this document does not impact existing SCTP messages on the wire. Also, this document does not introduce any new SCTP messages on the wire that require new security considerations.

SCTP-PF makes SCTP not only more robust during primary path failure/congestion, but also more vulnerable to network connectivity/congestion attacks on the primary path. SCTP-PF makes it easier for an attacker to trick SCTP into changing the data transfer path, since the duration of time that an attacker needs to negatively influence the network connectivity is much shorter than used in [RFC4960]. However, SCTP-PF does not constitute a significant change in the duration of time and effort an attacker needs to keep SCTP away from

the primary path. With the standard switchback operation in [RFC4960], SCTP resumes data transfer on its primary path as soon as the next HEARTBEAT succeeds.

On the other hand, usage of the Primary Path Switchover mechanism, does change the threat analysis. This is because on-path attackers can force a permanent change of the data transfer path by blocking the primary path until the switchover of the primary path is triggered by the Primary Path Switchover algorithm. This will especially be the case when the Primary Path Switchover is used together with SCTP-PF with the particular setting of PSMR = PFMR = 0, as Primary Path Switchover here happens already at the first RTO timeout experienced. Users of the Primary Path Switchover mechanism should be aware of this fact.

The event notification of path state transfer from active to PF state and vice versa gives attackers an increased possibility to generate more local events. However, it is assumed that event notifications are rate-limited in the implementation to address this threat.

9. MIB Considerations

SCTP-PF introduces new SCTP algorithms for failover and switchback with associated new state parameters. It is recommended that the SCTP-MIB defined in [RFC3873] is updated to support the management of the SCTP-PF implementation. This can be done by extending the sctpAssocRemAddrActive field of the SCTPAssocRemAddrTable to include information of the PF state of the destination address and by adding new fields to the SCTPAssocRemAddrTable supporting PotentiallyFailed.Max.Retrans (PFMR) and Primary.Switchover.Max.Retrans (PSMR) parameters.

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4960] Stewart, R., Ed., "Stream Control Transmission Protocol", RFC 4960, DOI 10.17487/RFC4960, September 2007, <http://www.rfc-editor.org/info/rfc4960>.

10.2. Informative References

[CARO02] Caro, A., Iyengar, J., Amer, P., Heinz, G., and R. Stewart, "A Two-level Threshold Recovery Mechanism for SCTP", Tech report, CIS Dept., University of Delaware, July 2002.

[CARO04] Caro, A., Amer, P., and R. Stewart, "End-to-End Failover Thresholds for Transport Layer Multihoming", MILCOM 2004, DOI 10.1109/MILCOM.2004.1493253, November 2004.

[CARO05] Caro, A., "End-to-End Fault Tolerance using Transport Layer Multihoming", Ph.D. Thesis, University of Delaware, DOI 10.1007/BF03219970, January 2005.

[FALLON08] Fallon, S., Jacob, P., Qiao, Y., Murphy, L., Fallon, E., and A. Hanley, "SCTP Switchover Performance Issues in WLAN Environments", IEEE CCNC, DOI 10.1109/ccnc08.2007.131, January 2008.

[GRINNEMO04] Grinnemo, K-J. and A. Brunstrom, "Performance of SCTP-controlled failovers in M3UA-based SIGTRAN networks", Advanced Simulation Technologies Conference, April 2004.

[IYENGAR06] Iyengar, J., Amer, P., and R. Stewart, "Concurrent Multipath Transfer using SCTP Multihoming over Independent End-to-end Paths", IEEE/ACM Transactions on Networking, DOI 10.1109/TNET.2006.882843, October 2006.

[JUNGMAIER02] Jungmaier, A., Rathgeb, E., and M. Tuexen, "On the use of SCTP in failover scenarios", World Multiconference on Systemics, Cybernetics and Informatics, July 2002.

[NATARAJAN09] Natarajan, P., Ekiz, N., Amer, P., and R. Stewart, "Concurrent Multipath Transfer during Path Failure", Computer Communications, DOI 10.1016/j.comcom.2009.05.001, May 2009.

[RFC3873] Pastor, J. and M. Belinchon, "Stream Control Transmission Protocol (SCTP) Management Information Base (MIB)", RFC 3873, DOI 10.17487/RFC3873, September 2004, <http://www.rfc-editor.org/info/rfc3873>.

[RFC6458] Stewart, R., Tuexen, M., Poon, K., Lei, P., and V. Yasevich, "Sockets API Extensions for the Stream Control Transmission Protocol (SCTP)", RFC 6458, DOI 10.17487/RFC6458, December 2011, <http://www.rfc-editor.org/info/rfc6458>.

Appendix A. Discussion of Alternative Approaches

This section lists alternative approaches for the issues described in this document. Although these approaches do not require updating RFC 4960, we do not recommend them for the reasons described below.

A.1. Reduce PMR

Smaller values for Path.Max.Retrans shorten the failover duration and in fact, this is recommended in some research results [JUNGMAIER02], [GRINNEMO04], and [FALLON08]. However, to significantly reduce the failover time, it is required to go down (as with PFMR) to Path.Max.Retrans=0 and, with this setting, SCTP switches to another destination address already on a single timeout that may result in spurious failover. Spurious failover is a problem in standard SCTP as the transmission of HEARTBEATs on the left primary path, unlike in SCTP-PF, is governed by HB.Interval also during the failover process. HB.Interval is usually set in the order of seconds (recommended value is 30 seconds) and when the primary path becomes inactive, the next HEARTBEAT may be transmitted only many seconds later: as recommended, only 30 seconds later. Meanwhile, the primary path may have long since recovered, if it needed recovery at all (indeed the failover could be truly spurious). In such situations, post failover, an endpoint is forced to wait in the order of many seconds before the endpoint can resume transmission on the primary path and furthermore, once it returns on the primary path, the CWND needs to be rebuilt anew -- a process that the throughput already had to suffer from on the alternate path. Using a smaller value for HB.Interval might help this situation, but it would result in a general waste of bandwidth as such more frequent HEARTBEATING would take place also when there are no observed troubles. The bandwidth overhead may be diminished by having the ULP use a smaller HB.Interval only on the path that, at any given time, is set to be the primary path; however, this adds complication in the ULP.

In addition, smaller Path.Max.Retrans values also affect the Association.Max.Retrans value. When the SCTP association’s error count exceeds Association.Max.Retrans threshold, the SCTP sender considers the peer endpoint unreachable and terminates the association. Section 8.2 in [RFC4960] recommends that the Association.Max.Retrans value should not be larger than the summation of the Path.Max.Retrans of each of the destination addresses.

Otherwise, the SCTP sender considers its peer reachable even when all destinations are INACTIVE. To avoid this dormant state operation, standard SCTP implementation SHOULD reduce Association.Max.Retrans accordingly whenever it reduces Path.Max.Retrans. However, smaller Association.Max.Retrans value decreases the fault tolerance of SCTP as it increases the chances of association termination during minor congestion events.

A.2. Adjust RTO-Related Parameters

As several research results indicate, we can also shorten the duration of the failover process by adjusting the RTO-related parameters [JUNGMAIER02] and [FALLON08]. During the failover process, RTO keeps being doubled. However, if we can choose a smaller value for RTO.max, we can stop the exponential growth of RTO at some point. Also, choosing smaller values for RTO.initial or RTO.min can contribute to keeping the RTO value small.

Similar to reducing Path.Max.Retrans, the advantage of this approach is that it requires no modification to the current specification, although it needs to ignore several recommendations described in Section 15 of [RFC4960]. However, this approach requires having enough knowledge about the network characteristics between endpoints. Otherwise, it can introduce adverse side effects such as spurious timeouts.

The significant issue with this approach, however, is that even if the RTO.max is lowered to an optimal low value, as long as the Path.Max.Retrans is kept at the recommended value from [RFC4960], the reduction of the RTO.max doesn’t reduce the failover time sufficiently enough to prevent severe performance degradation during failover.

Appendix B. Discussion of the Path-Bouncing Effect

The methods described in the document can accelerate the failover process. Hence, they might introduce a path-bouncing effect in which the sender keeps changing the data transmission path frequently. This sounds harmful to the data transfer; however, several research results indicate that there is no serious problem with SCTP in terms of the path-bouncing effect (see [CARO04] and [CARO05]).

There are two main reasons for this. First, SCTP is basically designed for multipath communication, which means SCTP maintains all path-related parameters (CWND, ssthresh, RTT, error count, etc.) per each destination address. These parameters cannot be affected by

path bouncing. In addition, when SCTP migrates the data transfer to another path, it starts with the minimal or the initial CWND. Hence, there is little chance for packet reordering or duplicating.

Second, even if all communication paths between the end nodes share the same bottleneck, the SCTP-PF results in a behavior already allowed by [RFC4960].

Appendix C. SCTP-PF for SCTP Single-Homed Operation

For a single-homed SCTP association, the only tangible effect of the activation of SCTP-PF operation is enhanced failure detection in terms of potential notification of the PF state of the sole destination address as well as, for idle associations, more rapid entering, and notification, of inactive state of the destination address and more rapid endpoint failure detection. It is believed that neither of these effects are harmful, provided adequate dormant state operation is implemented. Furthermore, it is believed that they may be particularly useful for applications that deploy multiple SCTP associations for load-balancing purposes. The early notification of the PF state may be used for preventive measures as the entering of the PF state can be used as a warning of potential congestion. Depending on the PMR value, the aggressive HEARTBEAT transmission in PF state may speed up the endpoint failure detection (exceed of AMR threshold on the sole path error counter) on idle associations in the case with a relatively large HB.Interval value compared to RTO (e.g., 30 seconds) is used.

Acknowledgments

The authors would like to acknowledge members of the IETF Transport Area Working Group (tsvwg) for continuing discussions on this document and insightful feedback, and we appreciate continuous encouragement and suggestions from the Chairs of the tsvwg. We especially wish to thank Michael Tuexen for his many invaluable comments and for his substantial supports with the making of the document.

Authors' Addresses

Yoshifumi Nishida
GE Global Research
2623 Camino Ramon
San Ramon, CA 94583
United States

Email: nishida@wide.ad.jp

Preethi Natarajan
Cisco Systems
510 McCarthy Blvd.
Milpitas, CA 95035
United States

Email: prenatar@cisco.com

Armando Caro
BBN Technologies
10 Moulton St.
Cambridge, MA 02138
United States

Email: acaro@bbn.com

Paul D. Amer
University of Delaware
Computer Science Department - 434 Smith Hall
Newark, DE 19716-2586
United States

Email: amer@udel.edu

Karen E. E. Nielsen
Ericsson
Kistavaegen 25
Stockholm 164 80
Sweden

Email: karen.nielsen@tieto.com